A saturation property for the spectral-Galerkin approximation of a Dirichlet problem in a square

A saturation property for the spectral-Galerkin approximation of a Dirichlet problem in a square

C. Canuto[*], R.H. Nochetto[†], R. Stevenson[‡], and M. Verani[§]

December 14, 2018

Abstract

Both practice and analysis of p-FEMs and adaptive hp-FEMs raise the question what increment in the current polynomial degree p guarantees a p-independent reduction of the Galerkin error. We answer this question for the p-FEM in the simplified context of homogeneous Dirichlet problems for the Poisson equation in the two dimensional unit square with polynomial data of degree p. We show that an increment proportional to p yields a p-robust error reduction and provide computational evidence that a constant increment does not.

1 Motivation and statement of the result

High order finite element methods (FEMs) can exhibit exponential convergence for elliptic problems with piecewise analytic data, and thus have become the methods of choice in computational science and engineering for such problems. The seminal work of Babuška and collaborators [1, 15, 16] has established the mathematical foundations for the a priori design of meshes and distribution of polynomial degrees, and proved exponential convergence for corner and edge singularities. In contrast, adaptive hp-FEMs

[*]Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy (claudio.canuto@polito.it)

[†]Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA (rhn@math.umd.edu)

[‡]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands (r.p.stevenson@uva.nl)

[§]MOX-Dipartimento di Matematica, Politecnico di Milano, P.zza Leonardo Da Vinci 32, I-20133 Milano, Italy (marco.verani@polimi.it)

hinge on a posteriori error estimators, which help determine whether it is more convenient to locally refine the mesh or increase the polynomial degree to improve the resolution. Although exponential convergence is observe experimentally, it has never been proved rigorously with the exception of [9].

Our adaptive *hp*-FEM of [9] hinges on a quasi-best *hp*-approximation module due to Binev [6], which in turn guarantees instance optimality and thus exponential convergence. As any other adaptive *hp*-FEM, ours also has a module to reduce the PDE error by a fixed fraction for piecewise polynomial data thereby avoiding data oscillation. In fact, denoting by $v \in V$ (a Hilbert space) the solution of the PDE, given an *hp*-partition $\mathcal{D}$, this module needs to find a refined *hp*-partition $\widetilde{\mathcal{D}}$ such that the Galerkin approximations $v_{\mathcal{D}}$ and $v_{\widetilde{\mathcal{D}}}$ from the corresponding finite element spaces $V_{\mathcal{D}}$ and $V_{\widetilde{\mathcal{D}}}$ satisfy, for a universal constant $0 < \alpha < 1$,

$$\|v - v_{\mathcal{D}}\|_V \leq \alpha \, \|v_{\widetilde{\mathcal{D}}} - v_{\mathcal{D}}\|_V. \tag{1.1}$$

Thanks to Galerkin orthogonality w.r.t. to the energy scalar product, this is indeed equivalent to the *global saturation property*

$$\|v - v_{\mathcal{D}}\|_V \lesssim \|v_{\widetilde{\mathcal{D}}} - v_{\mathcal{D}}\|_V. \tag{1.2}$$

The module to reduce the PDE error in [9] was based on *h*-refinements driven by the a posteriori estimator of Melenk and Wohlmuth [17]. Since this estimator is not *p*-robust, saturation uniformly in p could not be guaranteed. In [10] we turned to the more efficient *p*-refinements driven by the equilibrated flux residual estimator of Braess, Pillwein and Schöberl [7], and Ern and Vorahlík [13, 14], which lead to a *p*-robust constant α. We showed that the global norm of the residual $r(v_{\mathcal{D}}) \in V'$ of $v_{\mathcal{D}}$, satisfying $\|r(v_{\mathcal{D}})\|_{V'} \eqsim \|v - v_{\mathcal{D}}\|_V$, can be localized to stars around the nodes $a \in \mathcal{A}$ of $\mathcal{D}$, i.e.,

$$\|r(v_{\mathcal{D}})\|_{V'}^2 \eqsim \sum_{a \in \mathcal{A}} \|r(v_{\mathcal{D}})\|_{V_a'}^2,$$

where V_a is a local space on the star. We further exploited that the data is piecewise polynomial in $V_{\mathcal{D}}$ to infer (1.2) from the *local saturation property*

$$\|r(v_{\mathcal{D}})\|_{V_a'} \lesssim \|r(v_{\mathcal{D}})\|_{(V_a \cap V_{\widetilde{\mathcal{D}}})'}, \qquad \forall a \in \mathcal{A}. \tag{1.3}$$

We finally reduced the validity of this property to the validity of a similar one for three auxiliary functionals defined on the reference triangle, involving interior or boundary polynomial data depending on $V_{\mathcal{D}}$ (see [10, Theorem

6.3]). We presented overwhelming computational evidence in [10] supporting the fact that in order to obtain a p-robust constant in (1.3), the local polynomial degree p of $V_{\mathcal{D}}$ should be increased by an amount proportional to p.

In this paper we take up this question again in a further simplified setting and give a rigorous answer. We denote by Ω the reference element, which we assume to be the square $\Omega = I^2$ for $d = 2$, where $I = (-1, 1)$. For $p \geq 0$, let $\mathbb{P}_p(\Omega)$ denote the space of polynomials of total degree $\leq p$ restricted to Ω, and let $\mathcal{V}_p := \mathbb{P}_p(\Omega) \cap H^1_0(\Omega)$. Given $f \in \mathbb{P}_p(\Omega)$ we consider the functional $\ell = \ell(f) \in H^{-1}(\Omega)$ given by $v \mapsto \int_\Omega fv$ for all $v \in H^1_0(\Omega)$. The local saturation property (1.3) formulated in Ω for the functional ℓ thus reads

$$\|\ell\|_{H^{-1}(\Omega)} \lesssim \|\ell\|_{\mathcal{V}_q'} \tag{1.4}$$

for some $q > p$. We point out that ℓ is similar to one of the three functionals derived in [10]; we expect that our argument below extends to the remaining two functionals, but omit the details.

If we equip $H^1_0(\Omega)$ with the energy inner product $(\cdot, \cdot)_{H^1_0(\Omega)} := (\nabla\cdot, \nabla\cdot)_{L^2(\Omega)}$ and resulting norm $\|\cdot\|_{H^1_0(\Omega)}$, the Riesz representation $u = u(f) \in H^1_0(\Omega)$ of ℓ is the solution of the variational problem

$$(u, v)_{H^1_0(\Omega)} = (f, v)_{L^2(\Omega)} \qquad \forall v \in H^1_0(\Omega); \tag{1.5}$$

in turn, this is the weak form of the classical Poisson equation

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{1.6}$$

On the other hand, the Riesz representation of the functional ℓ restricted to $\mathcal{V}_q$ is the Galerkin projection of u onto $\mathcal{V}_q$, i.e., $u_q = u_q(f) \in \mathcal{V}_q$ satisfying

$$(u_q, v_q)_{H^1_0(\Omega)} = (f, v_q)_{L^2(\Omega)} \qquad \forall v \in \mathcal{V}_q \tag{1.7}$$

(note that $\mathcal{V}_q$ reduces to $\{0\}$ unless $q \geq 4$, which we assume in the following). With the notation just introduced, the desired inequality (1.4) is therefore equivalent to

$$\|u\|_{H^1_0(\Omega)} \lesssim \|u_q\|_{H^1_0(\Omega)}, \tag{1.8}$$

which is another expression of the saturation property. We aim at establishing the following rigorous result.

Theorem 1.1 (saturation property)**.** *There exists a constant $C > 0$ such that for all $\lambda > 1$, any mapping $p \mapsto q = q(p)$ satisfying $q(p) > \max(\lambda p, p+4)$ yields*

$$\|u\|_{H^1_0(\Omega)} \leq C \frac{\lambda}{\lambda - 1} \|u_q\|_{H^1_0(\Omega)} \qquad \textit{for all } p \geq 0 \textit{ and } f \in \mathbb{P}_p(\Omega). \tag{1.9}$$

Since most *hp*-FEMs in the literature perform p-enrichment upon adding a constant increment to p, typically 1 or 2, one may wonder whether the preceding sufficient condition on q is also necessary. We now investigate this question computationally upon defining

$$C_{p,q,r} := \max_{f \in \mathbb{P}_p(\Omega)} \frac{\|u_r\|_{H_0^1(\Omega)}}{\|u_q\|_{H_0^1(\Omega)}}$$

where $r \gg q$ is chosen computationally so that u_r is sufficiently close to u in $H_0^1(\Omega)$; note that this is not a hidden saturation assumption because the value of r is not predetermined but found once the number $C_{p,q,r}$ stabilizes. This calculation reduces to an eigenvalue problem, already used in [10], and leads to Figure 1 for $q = p + k$ with $k = 2, 4, 6, 10$. We thus realize that

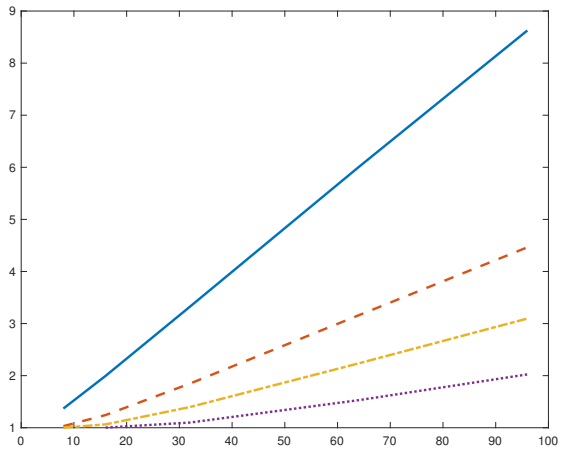

Figure 1: Constants $C_{p,q,r}$ vs $q = p + k$ for: $k = 2$ (solid), $k = 4$ (dashed), $k = 6$ (dashdot), $k = 10$ (dotted). The dependence is clearly linear rather than constant but the growth is moderate.

$C_{p,q,r}$ exhibits a modest but linear growth on $q = p+k$ for k constant, which confirms that this choice is not p-robust. For moderate values of p this might still be acceptable computationally, but it could compromise computational complexity for extreme values of p as in spectral algorithms [8].

Even though the saturation property is quite delicate, it has been often used in a posteriori error analysis of low order AFEMs until now. It originates in the work of Bank and Weiser [2], and Bornemann, Erdmann and

Kornhuber [3]; see Nochetto [18] for related work. Dörfler and Nochetto [12] proved the saturation property for $p = 1, q = 2$ and $d = 2$ provided data oscillation is small relative to $\|u - u_p\|_{H^1_0(\Omega)}$ but showed counterexamples for piecewise constant forcing f.

We stress that (1.9) is not asymptotic: it is valid for any $p \geq 0$ and any $f \in \mathbb{P}_p(\Omega)$. Since $u_q \to u$ in $H^1_0(\Omega)$ as $q \to \infty$ it is obvious that $C_{p,q,\infty} \to 1$ as $q \to \infty$. It is for this reason that Theorem 1.1 has some intrinsic value in the theory of FEMs and might have implications beyond a posteriori error analysis.

The proof of Theorem 1.1 proceeds as follows. We perform a multilevel decomposition of $\mathcal{V}_q$

$$\mathcal{V}_q = \bigoplus_{j=1}^{q} \mathcal{W}_j, \tag{1.10}$$

where $\mathcal{W}_j$ are polynomial subspaces of total degree j. Since this decomposition is quasi-orthogonal in the sense that

$$\mathcal{W}_j \perp \mathcal{W}_\ell \quad \text{for all } \ell \neq j-2, j, j+2,$$

we need to account for interactions between neighboring spaces $\mathcal{W}_j$. We study the angle between subspaces $\mathcal{W}_j$ and show it is larger than $\pi/3$; this is the content of Proposition 2.2. This in turn allows us to find the precise decay of high frequency modes of u_q, which leads to (1.9).

Such a rather technical derivation exploits the cartesian structure of the square Ω, which allows the use of a tensor-product modal basis in $\mathcal{V}_q$. Similar arguments work for the hypercube Ω for $d > 2$, at the expense of an increased complexity. If Ω is the reference simplex for $d = 2$, the Duffy transform maps the Koorwinder-Dubiner warped tensor-product basis in Ω to a subset of the tensor-product basis on the reference square examined in this paper. Therefore, the analysis below might also be relevant to derive the saturation property in the triangle. Overwhelming computational evidence reported in [10] indicates that such property is true, but a rigorous proof remains open.

The paper is organized as follows. In section 2 we introduce the multilevel decomposition (1.10) and discuss a few properties including Proposition 2.2. In section 3 we analyze the decay of high order components of u_q, whereas in section 5 we prove Theorem 1.1. We conclude in section 6 with the proof of the rather technical Proposition 2.2.

2 Multi-level decompositions of polynomial spaces

Hereafter, we recall the definition of classical polynomial bases in $L^2(\Omega)$ and in $H^1_0(\Omega)$, obtained by tensorization from corresponding bases in $L^2(I)$ and in $H^1_0(I)$, where $I = (-1, 1)$ is the reference interval. The elements of theses bases enjoy certain orthogonality properties, by which a multi-level, quasi-orthogonal decomposition of $H^1_0(\Omega)$ is obtained. This will be useful in deriving the main result of this paper.

On the interval I, we consider the *orthonormal Legendre basis in* $L^2(I)$

$$\vartheta_k(x) = \sqrt{k + \tfrac{1}{2}}\, L_k(x), \qquad k \geq 0, \tag{2.1}$$

(where L_k stands for the k-th Legendre orthogonal polynomial in I, which satisfies $\deg L_k = k$ and $L_k(1) = 1$), as well as the *orthonormal Babuška-Shen (BS) basis in* $H^1_0(I)$:

$$\begin{aligned} \varphi_k(x) &= \sqrt{k - \tfrac{1}{2}} \int_x^1 L_{k-1}(s)\, ds \\ &= \frac{1}{\sqrt{4k-2}} \big(L_{k-2}(x) - L_k(x)\big), \qquad k \geq 2\ . \end{aligned} \tag{2.2}$$

The BS basis enjoys the following orthogonality properties in $L^2(I)$ for $m \geq k$:

$$(\varphi_k, \varphi_m)_{L^2(I)} = \begin{cases} \frac{2}{(2k-3)(2k+1)} & \text{if } m = k\ , \\ -\frac{1}{(2k+1)\sqrt{(2k-1)(2k+3)}} & \text{if } m = k+2\ , \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

On the square $\Omega = I \times I$, the previous bases induce, resp., the *tensorized orthonormal Legendre basis in* $L^2(\Omega)$:

$$\Theta_k(x) = \vartheta_{k_1}(x_1)\vartheta_{k_2}(x_2), \qquad k \in \hat{\mathcal{K}}, \tag{2.4}$$

where $k = (k_1, k_2)$, $x = (x_1, x_2)$ and $\hat{\mathcal{K}} = \mathbb{N}^2$, and the *tensorized Babuška-Shen basis in* $H^1_0(\Omega)$:

$$\Phi_k(x) = \varphi_{k_1}(x_1)\varphi_{k_2}(x_2), \qquad k \in \mathcal{K}\ , \tag{2.5}$$

where $\mathcal{K} = \{k \in \mathbb{N}^2 \ :\ k_i \geq 2 \text{ for } i = 1, 2\}$.

The tensorized BS basis is not orthogonal in $H^1_0(\Omega)$. Indeed, from the expression

$$\begin{aligned} (\Phi_k, \Phi_m)_{H^1_0(\Omega)} = {} & (\varphi_{k_1}, \varphi_{m_1})_{H^1_0(I)} (\varphi_{k_2}, \varphi_{m_2})_{L^2(I)} \\ & + (\varphi_{k_1}, \varphi_{m_1})_{L^2(I)} (\varphi_{k_2}, \varphi_{m_2})_{H^1_0(I)}\ , \end{aligned}$$

and (2.3) we immediately obtain

$$(\Phi_k, \Phi_m)_{H^1_0(\Omega)} \neq 0 \ \text{ iff } \ \begin{cases} k_1 = m_1 \text{ and } k_2 - m_2 \in \{-2, 0, 2\}, \text{ or} \\ k_2 = m_2 \text{ and } k_1 - m_1 \in \{-2, 0, 2\}. \end{cases} \tag{2.6}$$

As a consequence, denoting by $|k| = |k_1| + |k_2|$ the ℓ^1-norm in $\mathbb{Z}^2$, we have

$$(\Phi_k, \Phi_m)_{H^1_0(\Omega)} = 0 \qquad \text{if } \ |k| - |m| \notin \{-2, 0, 2\}. \tag{2.7}$$

At last, concerning the interaction between the Legendre basis and the BS one, we have

$$(\Theta_k, \Phi_m)_{L^2(\Omega)} \neq 0 \ \text{ iff } \ k_1 \in \{m_1 - 2, m_1\} \text{ and } k_2 \in \{m_2 - 2, m_2\}, \tag{2.8}$$

which implies

$$(\Theta_k, \Phi_m)_{L^2(\Omega)} = 0 \qquad \text{if } \ |k - m| > 4. \tag{2.9}$$

Remark 2.1 (*orthogonality by parity).* Any function $v \in L^2(\Omega)$ can be split uniquely into four components

$$v = \sum_{\alpha \in \{0,1\}^2} v^\alpha, \tag{2.10}$$

where v^α for $\alpha = (\alpha_1, \alpha_2)$ is even (odd, resp.) with respect to the variable x_i ($i = 1, 2$) iff $\alpha_i = 0$ ($\alpha_i = 1$, resp.). For instance, $v^{(0,1)}$ satisfies $v^{(0,1)}(-x_1, x_2) = v^{(0,1)}(x_1, x_2)$ and $v^{(0,1)}(x_1, -x_2) = -v^{(0,1)}(x_1, x_2)$ for all $(x_1, x_2) \in \Omega$.

Components with different parity indices are always $L^2(\Omega)$-orthogonal, and $H^1_0(\Omega)$-orthogonal whenever $v \in H^1_0(\Omega)$. In particular, as a consequence of (2.6) and (2.8), we observe that $(\Phi_k, \Phi_m)_{H^1_0(\Omega)} = 0$ and $(\Theta_k, \Phi_m)_{L^2(\Omega)} = 0$ whenever k and m have at least one entry of different parity.

2.1 Detail spaces and their projectors

For $j \geq 4$, let us define the finite dimensional subspace of $H^1_0(\Omega)$

$$\mathcal{W}_j := \operatorname{span}\{\Phi_k \ : |k| = j\} \ . \tag{2.11}$$

Note that, thanks to (2.6), the functions Φ_k that generate $\mathcal{W}_j$ are mutually orthogonal in $H^1_0(\Omega)$. We immediately have the multi-level decompositions

$$\mathcal{V}_q = \bigoplus_{j=4}^{q} \mathcal{W}_j \quad \text{for all } q \geq 4 \qquad \text{and} \qquad H^1_0(\Omega) = \bigoplus_{j=4}^{\infty} \mathcal{W}_j \ . \tag{2.12}$$

Such decompositions are 'quasi-orthogonal', in the sense that by (2.7) we have

$$\mathcal{W}_j \perp_{H^1_0(\Omega)} \mathcal{W}_\ell \qquad \text{for all } \ \ell \neq j-2, j, j+2. \tag{2.13}$$

Furthermore, the 'angle' between two non-orthogonal subspaces is uniformly bounded away from 0, as implied by the following technical result, that will be crucial in the sequel. We postpone its proof to section 6.

Proposition 2.2 (angle between $\mathcal{W}_{j-2}$ and $\mathcal{W}_j$)**.** *Let $P_j : \mathcal{W}_{j-2} \to \mathcal{W}_j$ ($j \geq 6$) be the orthogonal projection with respect to the $H^1_0(\Omega)$-inner product. Then,*

$$\|P_j\|_{\mathcal{L}(\mathcal{W}_{j-2},\mathcal{W}_j)} < \frac{1}{2}.$$

Actually, there exists a constant $c > 0$ independent of j such that

$$\|P_j\|_{\mathcal{L}(\mathcal{W}_{j-2},\mathcal{W}_j)} \leq \frac{1}{2}\left(1 - \frac{c}{j^2}\right) .$$

Note that the orthogonal projection $P_j^* : \mathcal{W}_j \to \mathcal{W}_{j-2}$, given by the adjoint of P_j, satisfies the same estimate.

3 Decay of the higher-order components of the Galerkin solution

Given $f \in \mathbb{P}_p(\Omega)$, let $u_q \in \mathcal{V}_q$ be the Galerkin solution defined in (1.7), and let $u_q = \sum_{j=4}^{q} U_j$, with $U_j = U_j(q) \in \mathcal{W}_j$, be its multilevel decomposition according to (2.12). The purpose of this section is to prove that for any q sufficiently larger than p, the $H^1_0(\Omega)$-norm of U_q and U_{q-1} decay at least proportionally to the quantity $(q-p)^{-1}$. The precise result is as follows.

Proposition 3.1 (decay of U_j)**.** *For any $p \geq 0$ and $q > \hat{p} := p + 4$, one has*

$$\|U_j\|_{H^1_0(\Omega)} \leq \frac{6}{q-p} \|u_q\|_{H^1_0(\Omega)} , \qquad j = q,\, q-1. \tag{3.1}$$

Proof. We first observe that the parity splitting (2.10) of the forcing f induces by linearity a corresponding splitting of the Galerkin solution u_q as well as of each of its multi-level details U_j, which is nothing but the parity splitting of u_q as well as of U_j. Therefore, thanks to the orthogonality of the components with different parity (cf. Remark 2.1), it is enough to establish (3.1) for each component separately, and then sum-up the squares of both sides invoking Parseval's identity.

For the sake of definiteness, we will focus on the components of (even, even) type, the other types being amenable to a similar treatment. Thus, referring to (2.10) for the notation, we consider the component $u_q^{(0,0)}$ of u_q (which solves (1.7) for the forcing $f^{(0,0)}$), as well as its details $U_j^{(0,0)} \in \mathcal{W}_j^{(0,0)} := \text{span}\{\Phi_k : |k| = j \text{ and } k_1, k_2 \text{ are even}\}$. We aim at proving that for $q \geq \hat{p} + 1$ and $j \in \{q-1, q\}$

$$\|U_j^{(0,0)}\|_{H_0^1(\Omega)} \leq \frac{6}{q-p} \|u_q^{(0,0)}\|_{H_0^1(\Omega)}.$$

However, it is easily seen that $U_{q-1}^{(0,0)} = 0$ if q is even, and similarly $U_q^{(0,0)} = 0$ if q is odd. Hence, we will prove

$$\|U_q^{(0,0)}\|_{H_0^1(\Omega)} \leq \frac{6}{q-p} \|u_q^{(0,0)}\|_{H_0^1(\Omega)} \tag{3.2}$$

under the assumption that q is even, the other situation being similar.

To avoid cumbersome notation, for the rest of the proof we will drop the superscript $^{(0,0)}$ from all entities. So, we will write

$$u_q = \sum_{j=4}^{q}{}' U_j \qquad \text{with } U_j \in \mathcal{W}_j,$$

where here and in the sequel the symbol $'$ indicates that the summation runs over even indices only.

From the Galerkin equations, we have for any even $j \in [4, q]$

$$(u_q, W_j)_{H_0^1(\Omega)} = (f, W_j)_{L^2(\Omega)} \qquad \text{for all } W_j \in \mathcal{W}_j\ . \tag{3.3}$$

Since $q \geq \hat{p} + 1$, exploiting (2.7) and (2.8), (3.3) yields

$$(U_q, W_q)_{H_0^1(\Omega)} + (U_{q-2}, W_q)_{H_0^1(\Omega)} = 0 \qquad \text{for all } W_q \in \mathcal{W}_q\ , \tag{3.4}$$

which can be rewritten equivalently as

$$U_q = -T_q^{-1} P_q U_{q-2} \tag{3.5}$$

where $T_q = I$ and P_q is defined in Proposition 2.2.

For any even j satisfying $\hat{p} + 2 < j \leq q - 2$, (3.3) yields

$$(U_{j+2}, W_j)_{H_0^1(\Omega)} + (U_j, W_j)_{H_0^1(\Omega)} + (U_{j-2}, W_j)_{H_0^1(\Omega)} = 0 \tag{3.6}$$

for all $W_j \in \mathcal{W}_j$; this is equivalent to $P_{j+2}^* U_{j+2} + U_j + P_j U_{j-2} = 0$. Assuming by induction that $U_{j+2} = -T_{j+2}^{-1} P_{j+2} U_j$ with $\|T_{j+2}^{-1}\| \le 2$ (which is trivially true for $j = q-2$ according to (3.5)), gives

$$(I - P_{j+2}^* T_{j+2}^{-1} P_{j+2}) U_j = -P_j U_{j-2}.$$

Using Proposition 2.2, the operator

$$T_j := I - P_{j+2}^* T_{j+2}^{-1} P_{j+2} \tag{3.7}$$

is invertible and satisfies

$$\|T_j^{-1}\| \le \frac{1}{1 - \frac{1}{4}\|T_{j+2}^{-1}\|} \,. \tag{3.8}$$

We conclude that

$$U_j = -T_j^{-1} P_j U_{j-2}. \tag{3.9}$$

with $\|T_j^{-1}\| \le 2$, which proves the induction argument for all even j satisfying $\hat{p} + 2 < j \le q - 2$.

Next, we have to bound the norm of U_{j-2} for $j = \hat{p} + 4$ when p, hence $\hat{p}$, is even, or for $j = \hat{p} + 3$ when p is odd. It is therefore convenient to define the even integer

$$r := \begin{cases} \hat{p} = p + 4 & \text{if } p \text{ is even,} \\ \hat{p} - 1 = p + 3 & \text{if } p \text{ is odd,} \end{cases}$$

so that in both cases, we have to bound $\|U_{r+2}\|_{H_0^1(\Omega)}$. To this end, let us introduce

$$\mathcal{V}_r^{(0,0)} := \bigoplus_{j=4}^{r}{}' \mathcal{W}_j, \qquad \bar{u}_r := \sum_{j=4}^{r}{}' U_j \in \mathcal{V}_r^{(0,0)};$$

note that $\bar{u}_r \ne u_r$ because $U_j = U_j(q)$. Then, in view of (3.3), we deduce

$$(\bar{u}_r + \sum_{j=r+2}^{q}{}' U_j, v_r)_{H_0^1(\Omega)} = (f, v_r)_{L^2(\Omega)} \quad \text{for all } v_r \in \mathcal{V}_r^{(0,0)} \,, \tag{3.10}$$

which, thanks to $(\sum_{j=r+2}^{q}{}' U_j, v_r)_{H_0^1(\Omega)} = (P_r^* U_{r+2}, v_r)_{H_0^1(\Omega)}$ for all $v_r \in \mathcal{V}_r^{(0,0)}$, implies that

$$\bar{u}_r + P_r^* U_{r+2} = u_r.$$

We observe that (3.6) is also valid for $j = r+2$. Since $(U_r, W_{r+2})_{H_0^1(\Omega)} = (\bar{u}_r, W_{r+2})_{H_0^1(\Omega)}$ we obtain

$$(U_{r+4}, W_{r+2})_{H_0^1(\Omega)} + (U_{r+2}, W_{r+2})_{H_0^1(\Omega)} + (\bar{u}_r, W_{r+2})_{H_0^1(\Omega)} = 0 \qquad (3.11)$$

and

$$T_{r+2}U_{r+2} = -P_{r+2}\bar{u}_r$$

as it happened with (3.9). This implies

$$(T_{r+2} - P_{r+2}P_r^*)U_{r+2} = -P_{r+2}u_r \qquad (3.12)$$

which in view of (3.7) yields

$$(I - P_{r+4}^* T_{r+4}^{-1} P_{r+4} - P_{r+2}P_r^*)U_{r+2} = -P_{r+2}u_r. \qquad (3.13)$$

Since $\|P_{r+4}^* T_{r+4}^{-1} P_{r+4} + P_{r+2}P_r^*\| \le 2\frac{1}{4} + \frac{1}{4} = \frac{3}{4}$, thanks to Proposition 2.2 and $\|T_{r+4}^{-1}\| \le 2$, we conclude that

$$\|U_{r+2}\|_{H_0^1(\Omega)} \le \frac{1}{1-\frac{3}{4}}\frac{1}{2}\|u_r\|_{H_0^1(\Omega)} = 2\|u_r\|_{H_0^1(\Omega)} \le 2\|u_q\|_{H_0^1(\Omega)},$$

where last inequality follows from the inclusion $\mathcal{V}_r^{0,0)} \subset \mathcal{V}_q^{(0,0)}$ and the minimization property of the Galerkin solution.

Collecting the above results we arrive at

$$\begin{aligned} &U_j = -T_j^{-1}P_jU_{j-2} \qquad r+4 \le j \le q \quad (j \text{ even}),\\ &\|U_{r+2}\|_{H_0^1(\Omega)} \le 2\|u_q\|_{H_0^1(\Omega)}. \end{aligned} \qquad (3.14)$$

For $q \ge \hat{p}+4$, this implies

$$\|U_q\|_{H_0^1(\Omega)} \le \prod_{j=r+4}^{q}{}' \|T_j^{-1}P_j\| \|U_{r+2}\|_{H_0^1(\Omega)} \le 2\|u_q\|_{H_0^1(\Omega)} \prod_{j=r+4}^{q}{}' \|T_j^{-1}P_j\|.$$

In order to bound the product on the right-hand side, let us write $j = q-2m$ with $m = 0, 1, \ldots, s$ and $s := \frac{1}{2}(q-r) - 2$. Then, by Proposition 2.2, we have $\|T_j^{-1}P_j\| \le \frac{1}{2}\|T_j^{-1}\| =: \alpha_m$. Recalling (3.8), it holds

$$\alpha_m \le \frac{\frac{1}{2}}{1-\frac{1}{2}\alpha_{m-1}} = \frac{1}{2-\alpha_{m-1}}, \qquad \text{with } \alpha_0 \le \tfrac{1}{2}\,.$$

By recurrence, it is immediate to check that $\alpha_m \leq \frac{m+1}{m+2}$, whence

$$\prod_{j=r+4}^{q}{}' \|T_j^{-1} P_j\| \leq \prod_{m=0}^{s} \alpha_m \leq \prod_{m=0}^{s} \frac{m+1}{m+2} = \frac{1}{s+2} = \frac{2}{q-r}.$$

Since $q \geq p+6$ if p is even and $q \geq p+5$ if p is odd, it is easily checked that

$$\frac{2}{q-r} = \begin{cases} \frac{2}{(q-p)-4} \leq \frac{6}{q-p} & \text{if } p \text{ is even,} \\ \frac{2}{(q-p)-3} \leq \frac{5}{q-p} & \text{if } p \text{ is odd.} \end{cases}$$

This gives the desired estimate (3.2). □

4 A subspace decomposition in $H_0^1(\Omega)$

Consider the complementary space of $\mathcal{V}_q$ in $H_0^1(\Omega)$ given by

$$\mathcal{V}_q^c := \mathrm{clos}_{H_0^1(\Omega)} \,\mathrm{span}\,\{\Phi_m \,:\, |m| > q\}. \tag{4.1}$$

Therefore, $H_0^1(\Omega) = \mathcal{V}_q \oplus \mathcal{V}_q^c$ and any $v \in H_0^1(\Omega)$ can be split as

$$v = v_q + z_q, \qquad v_q \in \mathcal{V}_q, \quad z_q \in \mathcal{V}_q^c.$$

The purpose of this section is to apply once more Proposition 2.2 and derive a bound on the norm of v_q and z_q in terms of the norm of v.

We start with the following auxiliary result for any $w = \sum_{j=4}^{q} W_j \in \mathcal{V}_q$.

Lemma 4.1 (bound of $\|W_q\|_{H_0^1(\Omega)}$)**.** *For any $q \geq 4$ and any $w = \sum_{j=4}^{q} W_j \in \mathcal{V}_q$, one has*

$$\|W_q\|_{H_0^1(\Omega)} \leq \sqrt{2}\|w\|_{H_0^1(\Omega)} \,.$$

Proof. As in the previous section, splitting w and W_q in their orthogonal components according to the parity of the basis functions, it is enough to establish the result for each component separately. Hereafter, we detail the analysis for the '(even, even)' case, in which case we may assume q even, since otherwise $W_q^{(0,0)} = 0$ and the result is trivial.

Dropping as above the superscript $^{(0,0)}$ in functions and subspaces, we write $w = W + W_q$ with

$$W = \sum_{j=4}^{q-2}{}' W_j \in \mathcal{V}_{q-2}.$$

Keeping W_q fixed, let us first minimize the norm of w, i.e., let us look for the minimizer $\bar{W} \in \mathcal{V}_{q-2}$ of the quantity $\Psi(W) := \|W + W_q\|^2_{H^1_0(\Omega)}$. Such a function satisfies

$$(\bar{W}, Y)_{H^1_0(\Omega)} = -(W_q, Y)_{H^1_0(\Omega)} \qquad \text{for all } Y \in \mathcal{V}_{q-2} \tag{4.2}$$

and

$$\Psi(\bar{W}) = \|W_q\|^2_{H^1_0(\Omega)} + (W_q, \bar{W})_{H^1_0(\Omega)} \, . \tag{4.3}$$

Using the orthogonality conditions (2.7), we obtain the sequence of equations

$$(\bar{W}_4, Y_4)_{H^1_0(\Omega)} + (\bar{W}_6, Y_4)_{H^1_0(\Omega)} = 0 \qquad \text{for all } Y_4 \in \mathcal{W}_4 \, ,$$

and

$$(\bar{W}_{j-2}, Y_j)_{H^1_0(\Omega)} + (\bar{W}_j, Y_j)_{H^1_0(\Omega)} + (\bar{W}_{j+2}, Y_j)_{H^1_0(\Omega)} = 0 \qquad \text{for all } Y_j \in \mathcal{W}_j$$

for any even j such that $4 < j < q-2$, and finally

$$(\bar{W}_{q-4}, Y_{q-2})_{H^1_0(\Omega)} + (\bar{W}_{q-2}, Y_{q-2})_{H^1_0(\Omega)} = -(W_q, Y_{q-2})_{H^1_0(\Omega)}$$

for all $Y_{q-2} \in \mathcal{W}_{q-2}$. Setting recursively $T_4 = I$ and $T_j = (I - P_j T_{j-2}^{-1} P_j^*)$, we derive $\bar{W}_j = -T_j^{-1} P_{j+2}^* \bar{W}_{j+2}$ for $j = 4, 6, \ldots, q-4$, and $\bar{W}_{q-2} = -T_{q-2}^{-1} P_q^* W_q$. Note that, thanks to Proposition 2.2, one can prove as in Sect. 3 that $\|T_j^{-1}\| \le 2$ for all j. Since

$$\begin{aligned}(W_q, \bar{W})_{H^1_0(\Omega)} &= (W_q, \bar{W}_{q-2})_{H^1_0(\Omega)} \\ &= (P_q^* W_q, \bar{W}_{q-2})_{H^1_0(\Omega)} = -(P_q^* W_q, T_{q-2}^{-1} P_q^* W_q)_{H^1_0(\Omega)},\end{aligned}$$

using once more Proposition 2.2, we deduce

$$\begin{aligned}(P_q^* W_q, T_{q-2}^{-1} P_q^* W_q)_{H^1_0(\Omega)} &\le \|T_{q-2}^{-1}\| \, \|P_q^*\|^2 \|W_q\|^2_{H^1_0(\Omega)} \\ &\le 2 \left(\frac{1}{2}\right)^2 \|W_q\|^2_{H^1_0(\Omega)} = \frac{1}{2} \|W_q\|^2_{H^1_0(\Omega)} \, .\end{aligned}$$

In view of (4.3) and the preceding estimate, we conclude that

$$\|w\|^2_{H^1_0(\Omega)} = \Psi(W) \ge \Psi(\bar{W}) \ge \|W_q\|^2_{H^1_0(\Omega)} - \frac{1}{2} \|W_q\|^2_{H^1_0(\Omega)} = \frac{1}{2} \|W_q\|^2_{H^1_0(\Omega)}$$

for any $w \in \mathcal{V}_q^{(0,0)}$, whence the asserted estimate follows. □

We now establish the main result of this section.

Proposition 4.2 (control of $\|z_q\|_{H^1_0(\Omega)}$)**.** *There exists a constant $C_2 > 0$ such that for any $q \geq 4$ and any $v = v_q + z_q \in \mathcal{V}_q \oplus \mathcal{V}_q^c$, one has*

$$\|z_q\|_{H^1_0(\Omega)} \leq C_2\, q\, \|v\|_{H^1_0(\Omega)}. \tag{4.4}$$

Proof. Using $\|z_q\|_{H^1_0(\Omega)} \leq \|v\|_{H^1_0(\Omega)} + \|v_q\|_{H^1_0(\Omega)}$, it is enough to prove the existence of a constant $C_2' > 0$ independent of q such that for all $v \in H^1_0(\Omega)$

$$\|v_q\|_{H^1_0(\Omega)} \leq C_2' q \|v\|_{H^1_0(\Omega)}\ . \tag{4.5}$$

To this end, let us focus as above on the '(even, even)' components of v and v_q, in which case it is not restrictive to assume q even, and drop the superscript $^{(0,0)}$ in functions and subspaces. Let us fix any even integer $r > q$ and assume first that $v \in \mathcal{V}_r$ is written as $v = v_q + V$, with

$$V = \sum_{j=q+2}^{r}{}' V_j \in \bigoplus_{j=q+2}^{r}{}' \mathcal{W}_j.$$

By applying the same technique as above, i.e., minimizing the (squared) norm $\Psi(V) := \|v_q + V\|^2_{H^1_0(\Omega)}$, we find that

$$\|v\|^2_{H^1_0(\Omega)} = \Psi(V) \geq \Psi(\bar V) = \|v_q\|^2_{H^1_0(\Omega)} + (v_q, \bar V)_{H^1_0(\Omega)}\ , \tag{4.6}$$

where the minimizer $\bar V = \sum_{j=q+2}^{r}{}' \bar V_j$ is such that $\bar V_{q+2} = -T_{q+2}^{-1}\tilde P_{q+2} v_q$, for T_{q+2} defined recursively by (3.7) with $T_r = I$, and $\tilde P_{q+2} : H^1_0(\Omega) \to \mathcal{W}_{q+2}$ defined as the orthogonal projection in the $H^1_0(\Omega)$ inner product. Now,

$$\begin{aligned}(v_q, \bar V)_{H^1_0(\Omega)} &= (v_q, \bar V_{q+2})_{H^1_0(\Omega)} = (\tilde P_{q+2} v_q, \bar V_{q+2})_{H^1_0(\Omega)} \\ &= -(\tilde P_{q+2} v_q, T_{q+2}^{-1}\tilde P_{q+2} v_q)_{H^1_0(\Omega)}\end{aligned}$$

with

$$\big|(\tilde P_{q+2} v_q, T_{q+2}^{-1}\tilde P_{q+2} v_q)_{H^1_0(\Omega)}\big| \leq \|T_{q+2}^{-1}\|\, \|\tilde P_{q+2} v_q\|^2_{H^1_0(\Omega)} \leq 2\|\tilde P_{q+2} v_q\|^2_{H^1_0(\Omega)}\ .$$

Writing $v_q = \sum_{j=4}^{q}{}' V_j$, one has $\tilde P_{q+2} v_q = \tilde P_{q+2} V_q = P_{q+2} V_q$, whence by Proposition 2.2 with $\varepsilon_j = cj^{-2}$ and Lemma 4.1 we get

$$\begin{aligned}\|\tilde P_{q+2} v_q\|_{H^1_0(\Omega)} = \|P_{q+2} V_q\|_{H^1_0(\Omega)} &\leq \frac{1}{2}\,(1 - \varepsilon_{q+2})\, \|V_q\|_{H^1_0(\Omega)} \\ &\leq \frac{1}{\sqrt{2}}\,(1 - \varepsilon_{q+2})\, \|v_q\|_{H^1_0(\Omega)}\ ,\end{aligned}$$

which gives

$$\left|(\tilde{P}_{q+2}v_q, T_{q+2}^{-1}\tilde{P}_{q+2}v_q)_{H_0^1(\Omega)}\right| \le (1-\varepsilon_{q+2})^2 \|v_q\|^2_{H_0^1(\Omega)} .$$

Then, from (4.6) we obtain

$$\|v\|^2_{H_0^1(\Omega)} \ge \varepsilon_{q+2}(2-\varepsilon_{q+2})\|v_q\|^2_{H_0^1(\Omega)} ,$$

which immediately yields (4.5) for all $v \in \mathcal{V}_r = \mathcal{V}_r^{(0,0)}$ and all $r > q$.

The same result holds for all other combinations of parity indices; hence, it holds for any $v \in \mathcal{V}_r$. Since polynomials vanishing on $\partial\Omega$ form a dense subset of $H_0^1(\Omega)$, we conclude that (4.5) holds for all $v \in H_0^1(\Omega)$. □

5 Proof of Theorem 1.1

We actually prove the equivalent condition

$$\|u - u_q\|_{H_0^1(\Omega)} \lesssim \|u_q\|_{H_0^1(\Omega)},$$

and for that we write

$$\|u - u_q\|_{H_0^1(\Omega)} = \sup_{v\in H_0^1(\Omega),\ v\neq 0} \frac{(u-u_q, v)_{H_0^1(\Omega)}}{\|v\|_{H_0^1(\Omega)}}.$$

As in the previous section, let us split any $v \in H_0^1(\Omega)$ as $v = v_q + z_q \in \mathcal{V}_q \oplus \mathcal{V}_q^c$, where $\mathcal{V}_q^c$ is given by (4.1). By the Galerkin orthogonality and the definition of u, we have

$$(u-u_q, v)_{H_0^1(\Omega)} = (u-u_q, z_q)_{H_0^1(\Omega)} = (f, z_q)_{L^2(\Omega)} - (u_q, z_q)_{H_0^1(\Omega)}.$$

Recalling (2.8) and the condition $q > \hat{p}$, we have $(f, z_q)_{L^2(\Omega)} = 0$, hence

$$(u-u_q, v)_{H_0^1(\Omega)} = -(u_q, z_q)_{H_0^1(\Omega)}.$$

Now, recalling (2.12), we expand the Galerkin solution u_q as $u_q = \sum_{j=4}^{q} U_j$. Invoking (2.7), we get

$$(u_q, z_q)_{H_0^1(\Omega)} = (U_{q-1} + U_q, z_q)_{H_0^1(\Omega)}.$$

Applying Propositions 3.1 and 4.2, we get the following bound

$$(u-u_q, v)_{H_0^1(\Omega)} \le C_1 C_2 \frac{q}{q-p}\|u_q\|_{H_0^1(\Omega)}\|v\|_{H_0^1(\Omega)}.$$

Since for $q > \lambda p$, the relation $\frac{q}{q-p} < \frac{\lambda}{\lambda-1}$ holds, and the proof is complete.

6 Proof of Proposition 2.2

We establish the bound $\|P_{j+2}\|_{\mathcal{L}(\mathcal{W}_j,\mathcal{W}_{j+2})} \le \frac{1}{2}\big(\frac{1}{2} - \frac{c}{j^2}\big)$ for any $j \ge 4$. This will be achieved through various steps: we bound $\|P_{j+2}\|_{\mathcal{L}(\mathcal{W}_j,\mathcal{W}_{j+2})}$ in section 6.1 by the ℓ^∞-norm of a suitable matrix; in sections 6.2 and 6.3 we characterize such an ℓ^∞-norm and show that the desired bound reduces to certain properties of a suitable function; we finally analyze such function in section 6.4. This analysis is computer assisted.

6.1 Bounding the operator norm by a matrix norm

Recalling the definition (2.11) of the subspaces $\mathcal{W}_j$ as well as Remark 2.1, we can split each $\mathcal{W}_j$ into its two nontrivial orthogonal components according to parity; precisely, if j is even we have $\mathcal{W}_j = \mathcal{W}_j^{(0,0)} \oplus \mathcal{W}_j^{(1,1)}$, whereas if j is odd we have $\mathcal{W}_j = \mathcal{W}_j^{(1,0)} \oplus \mathcal{W}_j^{(0,1)}$. Furthermore, again by orthogonality it holds $P_{j+2} \in \mathcal{L}(\mathcal{W}_j^\alpha, \mathcal{W}_{j+2}^\alpha)$ for any $\alpha \in \{0,1\}^2$; hence, our target result can be achieved by considering each parity component separately. Hereafter, we will analyze the case j even and $\alpha = (0,0)$, i.e., we will bound $\|P_{j+2}\|_{\mathcal{L}(\mathcal{W}_j^{(0,0)},\mathcal{W}_{j+2}^{(0,0)})}$; the other three cases can be treated similarly.

Let us set $d_j := \dim \mathcal{W}_j^{(0,0)}$, note that $d_j = \frac{j}{2} - 1$ because $j = |h| = h_1 + h_2$ with $h_1, h_2 \ge 2$ even, and let us introduce the normalized basis functions $\hat\Phi_h := \Phi_h / \|\Phi_h\|_{H^1_0(\Omega)}$. For the sake of definiteness, let us order the basis functions in each $\mathcal{W}_j^{(0,0)}$ by increasing the first index h_1. Any $v \in \mathcal{W}_j^{(0,0)}$ is represented as

$$v = \sum_{|h|=j}{}' \hat v_h \hat\Phi_h \quad \text{with} \quad \mathbf{v} := (\hat v_h) \in \mathbb{R}^{d_j};$$

the summation symbol means that only indices $h = (h_1, h_2)$ with even components are considered. Similarly, any $w \in \mathcal{W}_{j+2}^{(0,0)}$ is represented as

$$w = \sum_{|k|=j+2}{}' \hat w_k \hat\Phi_k \quad \text{with} \quad \mathbf{w} := (\hat w_k) \in \mathbb{R}^{d_{j+2}}.$$

Therefore, if $w = P_{j+2}v$, then $\mathbf{w} = \mathbf{A}_j^T \mathbf{v}$, where $\mathbf{A}_j \in \mathbb{R}^{d_j \times d_{j+2}}$ is the matrix whose entries are

$$a_{hk} := (\hat\Phi_h, \hat\Phi_k)_{H^1_0(\Omega)} \quad \text{for} \quad |h| = j,\ |k| = j+2,$$

and h_1, h_2, k_1, k_2 even (recall that the Φ_h's that span $\mathcal{W}_j^{(0,0)}$ form an orthogonal basis for this space). Note that $\mathbf{A}_j$ is a sub-block of the (even, even) block $\mathbf{A}^{0,0}$ of the stiffness matrix $\mathbf{A}$ for the normalized Babuška-Shen basis in $H_0^1(\Omega)$. More precisely, denoting by $\mathbf{I}_j \in \mathbb{R}^{d_j \times d_j}$ the identity matrix of order d_j, we have $\mathbf{A}^{0,0} = \mathsf{tridiag}\ (\mathbf{A}_{j-2}^T,\ \mathbf{I}_j,\ \mathbf{A}_j)$ (with $j \geq 4$).

Now, one immediately has

$$\|P_{j+2}\|_{\mathcal{L}(\mathcal{W}_j^{(0,0)},\mathcal{W}_{j+2}^{(0,0)})} = \|\mathbf{A}_j^T\|_2 = \|\mathbf{A}_j\|_2 = \|P_{j+2}^*\|_{\mathcal{L}(\mathcal{W}_{j+2}^{(0,0)},\mathcal{W}_j^{(0,0)})}$$

(where $\|\cdot\|_p$ denotes the p-norm of a matrix), which together with the inequality $\|\mathbf{A}_j\|_2 = \rho(\mathbf{A}_j\mathbf{A}_j^T)^{1/2} \leq \|\mathbf{A}_j\mathbf{A}_j^T\|_\infty^{1/2}$, yields the bound

$$\|P_{j+2}\|_{\mathcal{L}(\mathcal{W}_j^{(0,0)},\mathcal{W}_{j+2}^{(0,0)})} \leq \|\mathbf{A}_j\mathbf{A}_j^T\|_\infty^{1/2}\ . \tag{6.1}$$

6.2 A first expression for the matrix entries

In order to compute the norm on the right-hand side of (6.1), we observe that $\mathbf{A}_j$ is a bi-diagonal matrix by condition (2.6). In fact, for any index h with $|h| = j$ the only indexes h', h'' with $|h'| = |h''| = j+2$ that give rise to entries $a_{h,h'}$ and $a_{h,h''}$ different from 0 are $h' = h+(0,2)$ and $h'' = h+(2,0)$. The explicit value of these entries is computable via the following formulas (in which all inner-products and norms are those of $H_0^1(\Omega)$):

$$a_{h,h'} = (\hat{\Phi}_h, \hat{\Phi}_{h'}) = \frac{(\Phi_h, \Phi_{h'})}{\|\Phi_h\|\,\|\Phi_{h'}\|} \qquad\qquad a_{h,h''} = (\hat{\Phi}_h, \hat{\Phi}_{h''}) = \frac{(\Phi_h, \Phi_{h''})}{\|\Phi_h\|\,\|\Phi_{h''}\|}$$

with

$$(\Phi_h, \Phi_{h'}) = -\frac{1}{(2h_2+1)\sqrt{(2h_2-1)(2h_2+3)}},$$
$$(\Phi_h, \Phi_{h''}) = -\frac{1}{(2h_1+1)\sqrt{(2h_1-1)(2h_1+3)}},$$

and

$$\|\Phi_h\|^2 = \frac{2}{(2h_1-3)(2h_1+1)} + \frac{2}{(2h_2-3)(2h_2+1)},$$

whence

$$\|\Phi_{h'}\|^2 = \frac{2}{(2h_1-3)(2h_1+1)} + \frac{2}{(2h_2+1)(2h_2+5)},$$
$$\|\Phi_{h''}\|^2 = \frac{2}{(2h_1+1)(2h_1+5)} + \frac{2}{(2h_2-3)(2h_2+1)}.$$

Since $2 \le h_1 \le j-2$ (h_1 even), it is convenient to set $n := \frac{j}{2}$ and $h_1 := 2i$, with $1 \le i \le n-1$; consequently, $h_2 := j - h_1 = 2(n-i)$. Substituting these expressions in the previous formulas, we obtain

$$(\Phi_h, \Phi_{h'}) = -\frac{1}{(4(n-i)+1)\sqrt{(4(n-i)-1)(4(n-i)+3)}} =: a_i$$

$$(\Phi_h, \Phi_{h''}) = -\frac{1}{(4i+1)\sqrt{(4i-1)(4i+3)}} =: b_i,$$

and

$$\|\Phi_h\|^2 = \frac{2}{(4i-3)(4i+1)} + \frac{2}{(4(n-i)-3)(4(n-i)+1)} =: \phi_i,$$

$$\|\Phi_{h'}\|^2 = \frac{2}{(4i-3)(4i+1)} + \frac{2}{(4(n-i)+1)(4(n-i)+5)} =: \psi_i,$$

$$\|\Phi_{h''}\|^2 = \frac{2}{(4i+1)(4i+5)} + \frac{2}{(4(n-i)-3)(4(n-i)+1)} =: \eta_i.$$

Note that $b_i = a_{n-i}$ and $\phi_i = \phi_{n-i}$, $\eta_i = \psi_{n-i}$. Hence, for $1 \le i \le n-1$,

$$(\mathbf{A}_j)_{ii} = \frac{a_i}{\sqrt{\phi_i \psi_i}} =: \delta_i, \qquad (\mathbf{A}_j)_{i,i+1} = \frac{b_i}{\sqrt{\phi_i \eta_i}} = \frac{a_{n-i}}{\sqrt{\phi_{n-i}\psi_{n-i}}} = \delta_{n-i},$$

i.e., $\mathbf{A}_j = \mathsf{bidiag}\ (\delta_i,\ \delta_{n-i})$. Consequently, the nonzero entries of the matrix $\mathbf{A}_j \mathbf{A}_j^T \in \mathbb{R}^{d_j \times d_j}$ are

$$(\mathbf{A}_j\mathbf{A}_j^T)_{i,i-1} = \delta_i \delta_{n-i+1}, \quad (\mathbf{A}_j\mathbf{A}_j^T)_{ii} = \delta_i^2 + \delta_{n-i}^2, \quad (\mathbf{A}_j\mathbf{A}_j^T)_{i,i+1} = \delta_{i+1}\delta_{n-i},$$

i.e., $\mathbf{A}_j\mathbf{A}_j^T = \mathsf{tridiag}\ (\delta_i\delta_{n-i+1},\ \delta_i^2 + \delta_{n-i}^2,\ \delta_{i+1}\delta_{n-i})$. Let us denote by $\mathsf{s}_i^{(j)}$ the sum of the entries in the i-th row of the matrix $\mathbf{A}_j\mathbf{A}_j^T$, which are all non-negative. Setting for convenience $\delta_n = 0$, we thus have

$$\mathsf{s}_i^{(j)} = \delta_i\delta_{n-i+1} + \delta_i^2 + \delta_{n-i}^2 + \delta_{i+1}\delta_{n-i}, \qquad 1 \le i \le n-1. \tag{6.2}$$

It is easily seen that $\mathsf{s}_i^{(j)} = \mathsf{s}_{n-i}^{(j)}$ for $1 \le i \le \frac{n}{2}$. Since

$$\|\mathbf{A}_j\mathbf{A}_j^T\|_\infty = \max_{1 \le i \le n-1} \mathsf{s}_i^{(j)}\ , \tag{6.3}$$

in view of (6.1) we are left with the problem of proving the existence of a constant $C > 0$ such that

$$\max_{1 \le i \le n-1} \mathsf{s}_i^{(j)} \le \frac{1}{4} - \frac{C}{j^2} \qquad \text{for all}\ \ j \ge 4; \tag{6.4}$$

indeed, thanks to $\sqrt{\frac{1}{4}-x} \le \frac{1}{2}-x$ for $x \le \frac{1}{4}$, we obtain Proposition 2.2 with $c = C$.

A direct computation shows that $\mathsf{s}_1^{(j)}$ and $\mathsf{s}_{n-1}^{(j)}$ satisfy the bound in (6.4) for a suitable C, because $\mathsf{s}_1^{(j)} = \mathsf{s}_{n-1}^{(j)} < \frac{1}{4}$ for all $j \ge 4$ and $\mathsf{s}_1^{(j)} \to \frac{3}{28}$ as $j \to \infty$. Thus, in the sequel we focus on the rows indexed from 2 to $n-2$, for $j \ge 8$ (i.e., $n \ge 4$).

6.3 A second expression for the matrix entries

We now apply a change of variables. Observing that all quantities a_i, ϕ_i, ψ_i, η_i, δ_i defined above depend upon $4i$ or $4(n-i)$ for $2 \le i \le n-2$, we first set $I := 4i$ and $N := 4n \ge 16$. To introduce the new variables (t, r), we first go back to the original range $1 \le i \le n-1$, i.e. $4 \le I \le N-4$, and parametrized I as follows

$$I = 4(1-t) + (N-4)t = 4 + Rt, \qquad 0 \le t \le 1,$$

with $R := N - 8 \ge 8$. Similarly, we write

$$N - I = 4 + R\tau, \qquad \tau = \tau(t) := 1 - t.$$

At last, we introduce the second parameter $r := \frac{1}{R} \le \frac{1}{8}$. With these notation at hand, we easily obtain the following expressions for a_i, ϕ_i and ψ_i:

$$a_i^2 = \frac{1}{R^4} \frac{1}{(\tau+3r)(\tau+5r)^2(\tau+7r)} =: \frac{1}{R^4} A(t,r),$$

$$\phi_i = \frac{1}{R^2} \left(\frac{2}{(t+r)(t+5r)} + \frac{2}{(\tau+r)(\tau+5r)} \right) =: \frac{1}{R^2} B(t,r),$$

$$\psi_i = \frac{1}{R^2} \left(\frac{2}{(t+r)(t+5r)} + \frac{2}{(\tau+5r)(\tau+9r)} \right) =: \frac{1}{R^2} C(t,r).$$

Hence, we arrive at

$$\delta_i^2 = \frac{a_i^2}{\phi_i \psi_i} = \frac{A(t,r)}{B(t,r)C(t,r)} =: D(t,r).$$

Straightforward computations show that

$$\delta_{i+1}^2 = D(t+4r, r), \qquad \delta_{n-i}^2 = D(\tau, r), \qquad \delta_{n-i+1}^2 = D(\tau+4r, r).$$

We conclude that the sum of the entries in the i-th row of $\mathbf{A}_j\mathbf{A}_j^T$, given by (6.2), can be expressed as follows:

$$\begin{aligned}\mathsf{s}_i^{(j)} &= \sqrt{D(t,r)D(\tau+4r,r)} + D(t,r) \\ &\quad + \sqrt{D(t+4r,r)D(\tau,r)} + D(\tau,r) =: S(t,r)\end{aligned} \tag{6.5}$$

for $2 \le i \le n-2$, which is equivalent to $4r \le t \le 1-4r$.

6.4 Bounding the matrix norm

Since the function $S(t,r)$ is symmetric with respect to $t = \frac{1}{2}$ for any r, we may restrict it to the triangle $0 \le t \le \frac{1}{2}$, $0 \le r \le \frac{1}{4}t$. Fig. 2 displays two plots of the function $\frac{1}{4} - S(t,r)$, and suggests clearly that $S(t,r) < \frac{1}{4}$ whenever $r > 0$, with a quadratic behavior in r at the origin. However, establishing such results rigorously is somehow complicated by the fact that $S(t,r)$ is singular at $(t,r) = (0,0)$, where it becomes multi-valued.

To remove this singularity, we apply the Duffy transform $(t,a) \mapsto (t,r) = (t,at)$, which maps the rectangle $0 \le t \le \frac{1}{2}$, $0 \le a \le \frac{1}{4}$ onto the triangle $0 \le t \le \frac{1}{2}$, $0 \le r \le \frac{1}{4}t$. Correspondingly, we are led to consider the function $\sigma(t,a) := S(t,at)$, which turns out to be smooth everywhere in this rectangle; a plot of the function $\frac{1}{4} - \sigma(t,a)$ is depicted in Fig. 3.

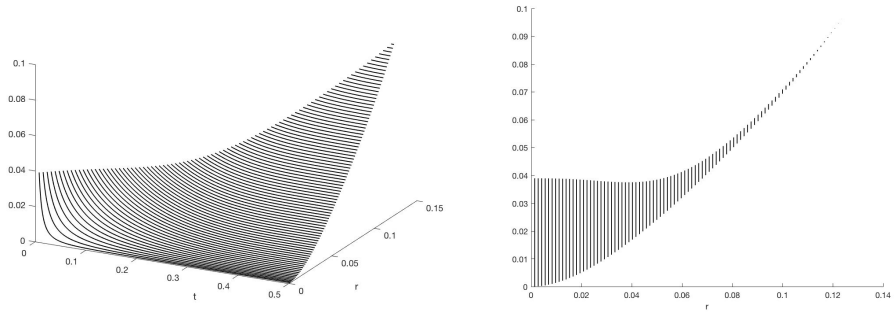

Figure 2: Two views of the graph of the function $\frac{1}{4} - S(t,r)$ on the triangle $0 \le t \le \frac{1}{2}$, $0 \le r \le \frac{1}{4}t$. Note that $S(t,r)$ is multi-valued at $t = r = 0$.

With the help of a symbolic manipulator, we obtain, for all $t \in [0, \frac{1}{2}]$,

$$\sigma(t,0) = \frac{1}{4}, \qquad \frac{\partial\sigma}{\partial a}(t,0) = 0, \qquad \frac{\partial^2\sigma}{\partial a^2}(t,0) = -\frac{G(t)}{\tau^2},$$

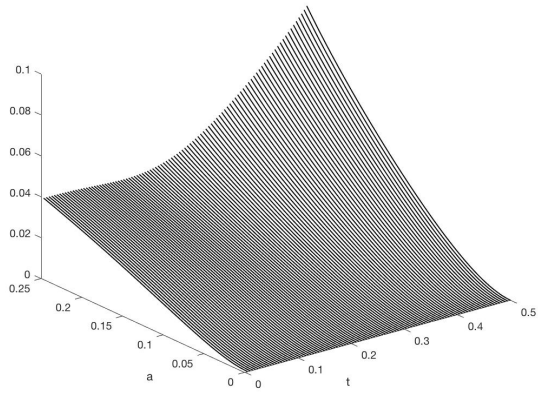

Figure 3: A view of the graph of the function $\frac{1}{4} - \sigma(t,a)$ in the rectangle $0 \le t \le \frac{1}{2}, 0 \le a \le \frac{1}{4}$.

where

$$G(t) := 3t^{10} + 9t^8\tau^2 - 8t^7\tau^3 + 16t^6\tau^4 + 24t^5\tau^5 + 16t^4\tau^6 - 8t^3\tau^7 + 9t^2\tau^8 + 3\tau^{10}.$$

We note that the polynomial $G(t)$ is strictly decreasing in $[0, \frac{1}{2}]$ between $G(0) = 3$ and $G(\frac{1}{2}) = 1$. We thus easily see that $\frac{\partial^2\sigma}{\partial a^2}(t,0) \le -1$ for $0 \le t \le \frac{1}{2}$; hence, by continuity we get the existence of two constants $C_* > 0$ and $a_* \in (0, \frac{1}{4}]$ such that $\frac{\partial^2\sigma}{\partial a^2}(t,a) \le -C_*$ for $0 \le t \le \frac{1}{2}$ and $0 \le a \le a_*$. With these constants at hand, by Taylor's expansion with Lagrange's reminder, we are entitled to write

$$\sigma(t,a) \;=\; \frac{1}{4} + \frac{\partial^2\sigma}{\partial a^2}(t,\bar{a})a^2 \;\le\; \frac{1}{4} - C_*a^2 \qquad \text{for } \; 0 \le t \le \tfrac{1}{2},\; 0 < a \le a_*,$$

with some $\bar{a} = \bar{a}(t,a) \in (0,a)$.

By computing the symbolic expression of the function $\frac{\partial^2\sigma}{\partial a^2}(t,a)$ and by examining its level sets (via a numerical procedure), one finds that a_* safely satisfies $a_* > \frac{1}{10}$ (see Figure 4). Therefore, going back to our function $S(t,r) = \sigma(t, \frac{r}{t})$, we deduce that

$$S(t,r) \le \frac{1}{4} - \frac{C_*}{t^2}r^2 \le \frac{1}{4} - 4C_*r^2 \qquad \text{for } \; 0 < r \le \tfrac{1}{10}t, \;\; t \le \tfrac{1}{2}.$$

Recalling (6.5) and using the expressions $t = 4r(i-1)$ and $r = \frac{1}{2}\frac{1}{j-4} > \frac{1}{2j}$,

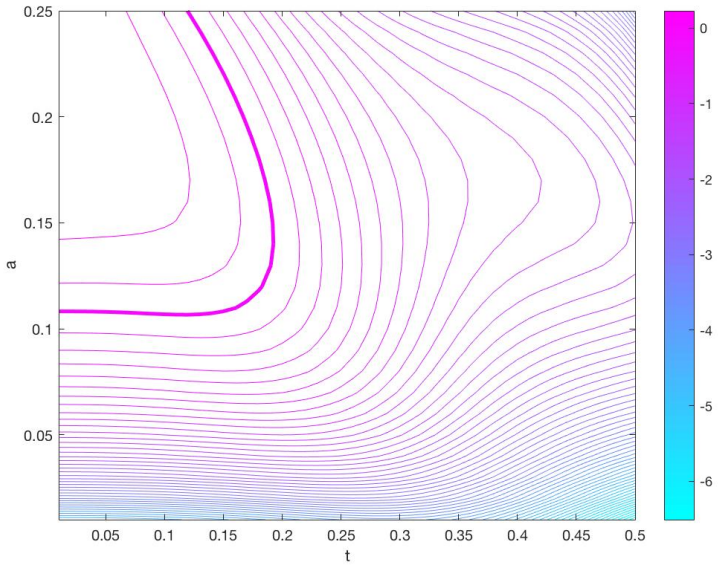

Figure 4: Contour plot of $\frac{\partial^2\sigma}{\partial a^2}(t,a)$: the function is negative on $[0,\frac{1}{2}]\times[0,a^*]$ with $a^* > \frac{1}{10}$. The thicker curve represents the zero level set

we immediately obtain

$$\mathsf{s}_i^{(j)} \le \frac{1}{4} - \frac{C_*}{j^2} \qquad \text{for} \quad 4 \le i \le \tfrac{n}{2}. \tag{6.6}$$

Note that we require the restriction $i \ge 4$ to satisfy the constraint

$$r \le \frac{1}{10}t = \frac{2}{5}r(i-1) \qquad \Rightarrow \qquad i \ge \frac{7}{2}.$$

Therefore, we are left with the task of establishing a similar bound for $\mathsf{s}_2^{(j)}$ and $\mathsf{s}_3^{(j)}$ by different means. It is easily checked that for $j \to \infty$ it holds $\mathsf{s}_2^{(j)} \to \frac{65}{308} < \frac{1}{4}$ and $\mathsf{s}_3^{(j)} \to \frac{55}{220} < \frac{1}{4}$, while both $\mathsf{s}_2^{(j)}$ and $\mathsf{s}_2^{(j)}$ are $< \frac{1}{4}$ for all $j \ge 8$. This implies the desired bound for a suitable constant $C_{**} > 0$.

The proof of (6.4) is thus complete, whence Proposition 2.2 is established.

Acknowledgements. C.C. carried out this work within the "Progetto di Eccellenza 2018-2022", granted by MIUR (Italian Ministry of University

and Research) to the Department of Mathematical Sciences, Politecnico di Torino.

The first and fourth authors are members of the INdAM research group GNCS, which granted partial support to this research.

References

[1] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta. Efficient preconditioning for the p-version finite element method in two dimensions. SIAM J. Numer. Anal., 28(3):624–661, 1991.

[2] R.E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations Math. Comp., 44:285–301, 1985.

[3] F. A. Bornemann, B. Erdmann, and R. Kornhuber, A posteriori error estimates for elliptic problems in two and three space dimensions, SIAM J. Numer. Anal., 33:1188–1204, 1996.

[4] M. Bürg and W. Dörfler. Convergence of an adaptive *hp* finite element strategy in higher space-dimensions. Appl. Numer. Math., 61(11):1132–1146, 2011.

[5] P. Binev, W. Dahmen, and R. DeVore. Adaptive finite element methods with convergence rates. Numer. Math., 97(2):219 – 268, 2004.

[6] P. Binev. Tree approximation for *hp*-adaptivity. SIAM J. Numer. Anal., 56(6):3346–3357, 2018

[7] D. Braess, V. Pillwein, and J. Schöberl. Equilibrated residual error estimates are *p*-robust. Comput. Methods Appl. Mech. Engrg., 198(13-14):1189–1197, 2009.

[8] C. Canuto, M. Y.Hussaini, A. Quarteroni, and T. A. Zang. *Spectral Methods. Fundamentals in Single Domains.* Springer 2016.

[9] C. Canuto, R.H. Nochetto, R. Stevenson, and M. Verani. Convergence and optimality of *hp*-AFEM. Numer. Math. 135 (2017), 1073-1119

[10] C. Canuto, R.H. Nochetto, R. Stevenson, and M. Verani. On *p*-robust saturation for *hp*-AFEM. Comput. & Math. with Appl. 73 (2017), 2004–2022.

[11] L. Demkowicz, J. Gopalakrishnan, and J. Schöberl. Polynomial extension operators. Part III. Math. Comp., 81(279):1289–1326, 2012.

[12] W. Dörfler and R.H. Nochetto, Small data oscillation implies the saturation assumption. Numer. Math. 91:1–12, 2002.

[13] A. Ern and M. Vohralík. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. SIAM J. Numer. Anal., 53(2):1058–1081, 2015.

[14] A. Ern and M. Vohralík. Stable broken H1 and H(div) polynomial extensions for polynomial-degree-robust potential and flux reconstruction in three space dimensions. hal 01422204, Inria Paris-Rocquencourt, 2016.

[15] B. Guo and I. Babuška. The *hp* version of finite element method, Part 1: The basic approximation results. Comp. Mech, 1:21–41, 1986.

[16] B. Guo and I. Babuška. The *hp* version of finite element method, Part 2: General results and application. Comp. Mech, 1:203–220, 1986.

[17] J. M. Melenk and B. I. Wohlmuth. On residual-based a posteriori error estimation in *hp*-FEM. *Adv. Comput. Math.*, 15(1-4):311–331, 2002.

[18] R.H. Nochetto, Removing the saturation assumption in a posteriori error analysis, Istit. Lombardo Accad. Sci. Lett. Rend. A, 127:67-82, 1993.